

## Statistical analysis of soy cultivation in Argentina

M. C. Miranda del Fresno<sup>1</sup>, Y. Villacampa<sup>2</sup>,  
F. J. Navarro-Gonzalez<sup>2</sup> & P. Sastre-Vázquez<sup>3</sup>

<sup>1</sup>*COCINET, Center for Research and Environmental Studies,  
University Nacional del Centro de la Provincia de Buenos Aires,  
Argentina*

<sup>2</sup>*Department of Applied Mathematics, University of Alicante, Spain*

<sup>3</sup>*Area of Mathematics, Faculty of Agriculture, University Nacional del  
Centro de la Provincia de Buenos Aires, Argentina*

### Abstract

In this paper, the different variables that characterize the soybean crop in Argentina are analyzed. The main objective is to obtain information that will be useful for the analysis of the environmental consequences of the progress of this crop in the country.

This has been made by a descriptive statistical analysis and a factor analysis was used to identify factors which characterize the soybean crop and also a classification or clustering between provinces with similar characteristics. The information parameters studied were: the tendency and dispersion measures for soybean yield of 15 provinces in which it was grown between the 2001/02 and 2010/11 or from 2001/02 to 2010/11; the correlations between area harvested acreage, seeded area, production and performance, and the factors that bind some variables that characterize the soybean crop that allows a grouping between provinces with similar characteristics. Analyses were performed by using data obtained from: 1) MinAgri (last 10 seasons) b) 2001 National Census c) CNA 2002. From a group of variables: region extension (ha); population density 2001; population of the region in 2001; NBI 2001, literate people older than 10 years 2001; illiterate people older than 10 years 2001; permanent workers in the farming sector 2002; seeded area (ha) 2002; cultivated area of oily plants (ha) 2002; tractors 2002; harvesters 2002; seeded area (ha) 2001/02; harvested area (ha) 2001/02; production (Tn) 2001/02.



Furthermore, using linear regressions, the relationships between: a) cultivated and harvested areas; b) total soy production according to the harvested area; and c) estimations of the productions over cultivated areas are given.

*Keywords: soy, Argentina, statistical analysis, seeded area, harvested area.*

## 1 Introduction

Argentina, with a huge territory, contains several kinds of climatic zones from cold to tropical climates, and from desert to subtropical regions. It is the seventh biggest country in the World, and most part of its territory locates on subtropical warm and wet zones. The presence of loess from Pleistocene has originated rich soils, which together with these favourable climates, have made Pampean region (see Figure 1) one of the six most important agricultural productive zones in the world, as it is sometimes called “the world’s breadbasket” [1].

The farming techniques were conservationist at the beginning, but today they are not. So, the increase of soy cultivation has caused the occupation of seeding spaces where less intensive techniques were used in the past (like rotations). Part of this process has been encouraged by the industries which use soy as prime factor to obtain oil and flours to feed animals [1]. In the 69/70 data from the Ministry of the area show that 30.470 ha were seeded, while by the year 2011/12, this area grows to 18.670.937, what is an increase rate of 61.276% in 42 years [2]. The association between glyphosate resistant transgenic soy, (glyphosate is a broad spectrum herbicide with high toxicity) of the *Glycine max* variety, and the use of direct seeding stimulated the advance of the farm frontier over native forest and zones of marginal interest from the farming point of view [3]. The production model is a success from the economic point of view, but has a set of negative social, environmental and sanitary side effects.

Different works describe some of their impacts: reduction in the number of medium and small producers, and the amount of rural workers; a big number of complaints about contamination caused by glyphosate (used in fumigations); loss of biodiversity; high dependency on transnational seed-provider firms [4]; the diminishing area of other crops and natural ecosystems, when they are replaced by soy; risks of ground erosion and water pollution [5]; deforestation; reduction in the variety of ground uses; the increasing of agrochemical substances use; etc. [6].

The expansion and importance of soy crop in Argentina, and the actual discussion about social and environmental costs of that soy expansion in Argentina, show the necessity of evaluating the effects of changes in the ground uses. These changes can originate social and economical conflicts. Changes can affect seriously to farm settlements developed without any planning and final consequences could affect to the entire society. In these cases a right knowledge of the dynamic of the system change can, through planning, give predictive answers avoiding losses and environmental damages.

Moreover, the description of change patterns, their special and temporal distributions and their relationships with other descriptive variables are important elements to do hypothesis on the process of farm expansion in a region. In



summary, it is very important to generate scientific knowledge on the consequences of human transformations on the territory, from social, economical and environmental scopes.

This work tries to describe the most important patterns in the change of cultivated and harvested areas, and the efficiency and production of soy in the regions of Argentina, in a period of ten years. Also, from the variations in the total cultivated areas and the fraction of soy cultivated areas, a geographical study of the distribution of the cultivation on the considered regions is done. Some socioeconomic variables are considered, grouping data from regions with similar characteristics.



Figure 1: Relative situation, analyzed regions, Pampa region and climates of Argentina.

## 2 Methodology

In this paper the most important patterns of change in the variables related to soy are described: cultivated surface, seeded surface, production and efficiency; the factors that relate some of these variables are identified and data from regions with similar characteristics are grouped. The data are obtained from: 1) Ministry



of Agriculture, Livestock and Fisheries (MinAgri) (last ten years) b) National Census 2001 and c) National Agricultural Census (CNA) 2002.

Estimations were made on the tendency and dispersion measures for efficiency of soy, the variable that reflects the amount of soy tonnes obtained by crop hectare. To establish the existence of a dependence relationship, the correlation between the variables harvested area, cultivated area, production and efficiency, was calculated, using the Pearson method. To be able to answer to questions as: How does soy production change when harvested area grows?, Which is the increase rate of soy efficiency as regards to cultivated area?, Which is the relationship between the harvested and seeded areas?, Which is the estimated percentage of seeded areas that are not harvested?, linear regressions were obtained, using: 1) Efficiency depending on harvested area and 2) Harvested area depending on seeded area.

To reduce the number of variables used, a factorial analysis was applied, by extracting the principal components, using the Varimax normalization with Kaiser as rotation method. With the obtained data, a clustering analysis was done. The factorial analysis was applied on the original variables and a new group of variables is selected: region extension (ha), population of the region in 2001, population density in 2001, needs unsatisfied basic (NBI) 2001, literate order than 10 years; illiterate older than 10 years 2001; permanent workers in farming sector 2002; seeded area 2002 (ha); cultivated area of oily plants (ha) 2002; tractors 2002; harvesters 2002; seeded area (ha) 2001/02; harvested area (ha) 2001/02; production (Tn) 2001/02. Finally, from the hierarchical clustering, a dendogram is obtained.

### 3 Results

From the data displayed in Table 1 and Figure 1, which show the descriptive analysis results for the efficiency data, it can be stated that:

Table 1: Descriptive analysis of efficiency (kg/ha).

Years	Efficiency									
	2001/02	2002/03	2003/04	2004/05	2005/06	2006/07	2007/08	2008/09	2009/10	2010/11
Average	2474,33	2376,93	2005,00	2156,60	2286,80	2549,40	2465,20	1773,80	2547,73	2388,40
Median	2430,00	2272,00	1950,00	2235,00	2421,00	2692,00	2457,00	1500,00	2602,00	2431,00
Std. Dev.	299,354	407,288	458,391	743,174	630,348	493,907	526,831	584,490	599,310	492,008
Minimum	2043	1800	1143	527	950	1768	1607	1076	1500	1500
Maximum	2900	3136	2900	3113	3003	3290	3349	2604	3649	3148



1. Averages have varied, having years with an increase (between years 2003/04 and 2006/07 and 2008/09 and 2009/10) and others with a decrease (between years 2001/02 and 2003/04; 2006/07 and 2008/09; 2009/10 and 2010/11).
2. The marginal averages show that the greatest efficiency was in the region of Santa Fe, and followed by, Buenos Aires, Córdoba, Catamarca, Misiones, Salta, Jujuy, Entre Ríos, and the others regions.
3. Median has moved under and over the average value.
4. Typical deviation has varied between 299,354 and 743,174, and the variance between 89.612,810 and 55.2307,400.
5. Minimum values vary from 527 to 2043; while maximum do between 2604 and 3649.

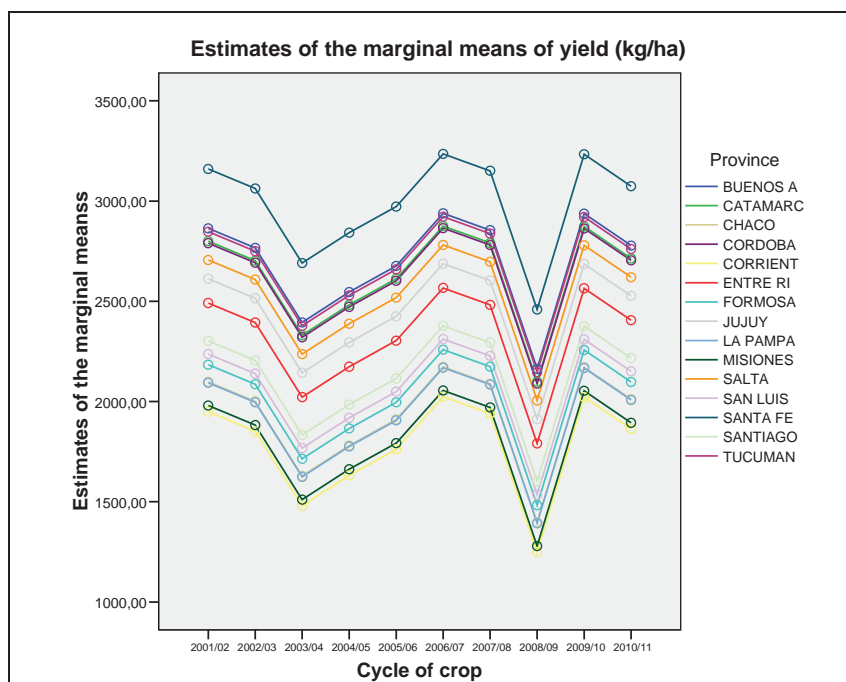


Figure 2: Estimated marginal averages of efficiency (kg/ha).

An analysis of the Pearson Correlation, between the variables analyzed in campaign 2001/02 and 2010/11, shows that: 1) There is a high correlation between the variables seeded area and harvested area, between production and seeded area and between production and harvested area. 2) There is no statistical significant correlation between efficiency and the other three variables. But, in campaign 2010/11 the association between the efficiency and the areas diminish to 0.31, while it was 0.33 in the campaign 2001/02. Both campaigns have the same association rate (0.36) between the performance and the production.

In Figure 3 all the obtained regressions are shown, with each R<sup>2</sup> value. The regression coefficients were, in all the cases, statistically significant.



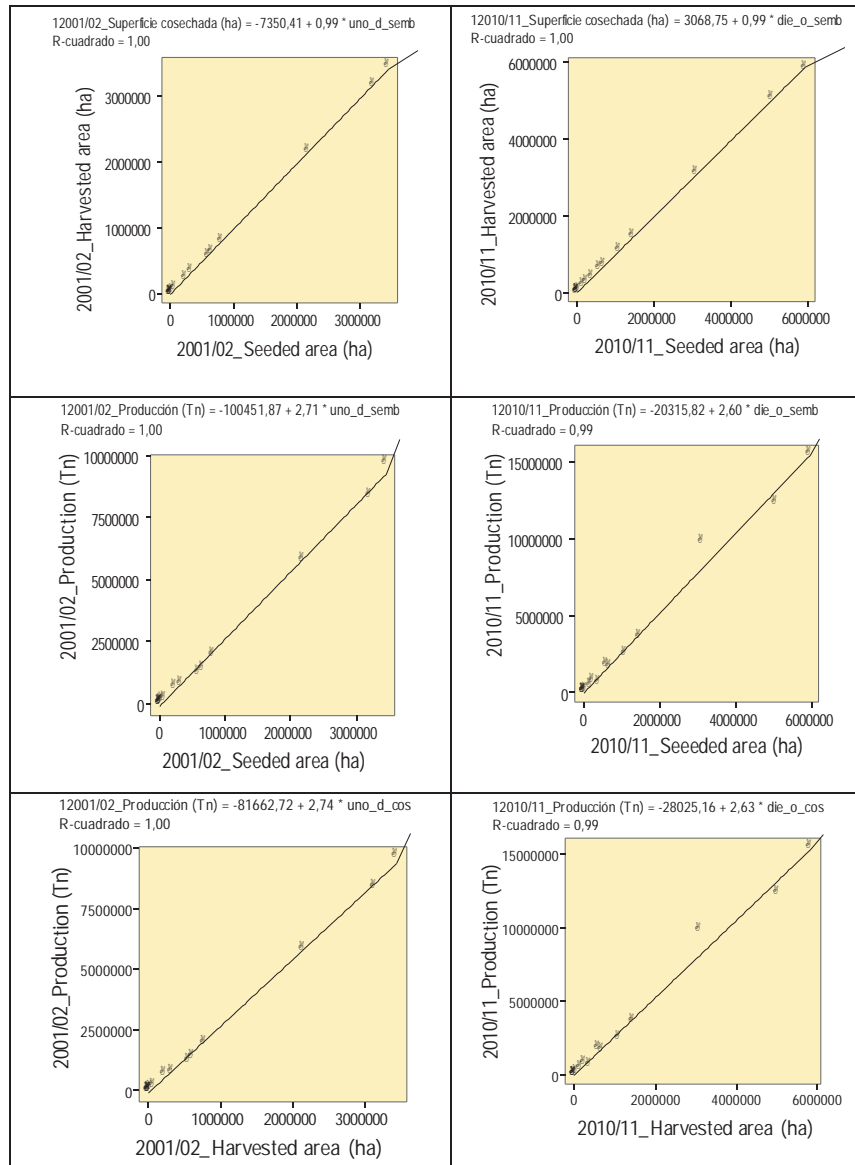


Figure 3: Linear regressions on years 2001/02 and 2010/11.

Using linear regressions, the relationships between cultivated and harvested areas are given by:

$$\text{Harvested area 2001/02} = -7350,41 + 0,99 (\text{Seeded area 2001/02})$$

$$\text{Harvested area 2001/02} = 3068,75 + 0,99 (\text{Seeded area 2010/11})$$



Total soy production as function of the harvested area in campaign 2001/02 is estimated as:

$$\text{Production 2001/02} = -81662,715 + 2,738 (\text{Harvested area 2002/01})$$

In the campaign 2010/11, the equation is:

$$\text{Production 2001/02} = -28025,164 + 2,630 \text{ Harvested area 2010/11}$$

This result shows that the increase rate of the production depending on harvested area has reduced from 2,7 (2001/02) to 2,6 (in 2010/11). Estimations for the productions over seeded areas in these campaigns are given by the equations:

$$\text{Production 2001/02} = -100451,87 + 2,71 \text{ Seeded area 2002/01}$$

$$\text{Production 2010/11} = -20315,82 + 2,60 \text{ Seeded area 2010/11}$$

The explained total variance table obtained from the principal component method, shows that the variables have been grouped using 2 factors, the first explains 77,4%, while the second gives account of 13,1%. So, the two main factors give a rate of 90,5%. The matrix of rotated components, Table 3, shows that the first factor, population component, it is easily explained by: NBI\_2001, population of the region in 2001; literate older than 10 years 2001; illiterate older than 10 years 2001; permanent workers in farming sector 2002; region extension (ha); population density 2001; and the second factor (productive component) can be explained mainly by the variables: 2001/02 production (Tn), 2001/02 harvested area (ha), 2001/02 seeded area(ha), cultivated area of oily plants 2002 (ha), tractors 2002, planted area 2002 (ha), harvesters 2002.

Table 2: Rotated component matrix.

Variable	Component	
	1	2
NBI_2001	,966	,246
Population of the region in 2001	,936	,334
Literate older than 10 years 2001	,933	,341
Illiterate older than 10 years 2001	,922	,317
Permanent workers in farming sector 2001	,764	,547
Region extension (ha)	,680	,464
Population density 2001	,614	
Production (Tn) 2001/02	,234	,967
Harvested area (ha) 2001/02	,239	,966
Seeded area (ha) 2001/02	,241	,964
Cultivated area of oily plants 2002 (ha)	,467	,878
Tractors 2002	,669	,730
Planted area 2002 (ha)	,654	,721
Harvesters 2002	,655	,704

The variance analysis (ANOVA), gives a significance of 0,00, that is, there is a big difference between the factors.



Using the last two factors: population and productive, a clustering analysis was made. From this, can be stated that: 1) Group 1 contains only the region of Buenos Aires; 2) The second group is composed of Catamarca, Formosa, Jujuy, La Pampa, Salta and San Luis; 3) The elements of the third group are Chaco, Entre Ríos and Santiago del Estero; 4) The fourth group contains Córdoba and Santa Fe; and 5) The last group is formed by Corrientes, Tucumán and Misiones.

Figure 4 shows the dendrogram for the population and productive factors, where at distance 1 just eight of the fifteen regions join in binary groups. At distance 2, these join any other, and twelve of them form a cluster. Out of them remain Córdoba and Santa Fe, and Buenos Aires that it is aside from the rest.

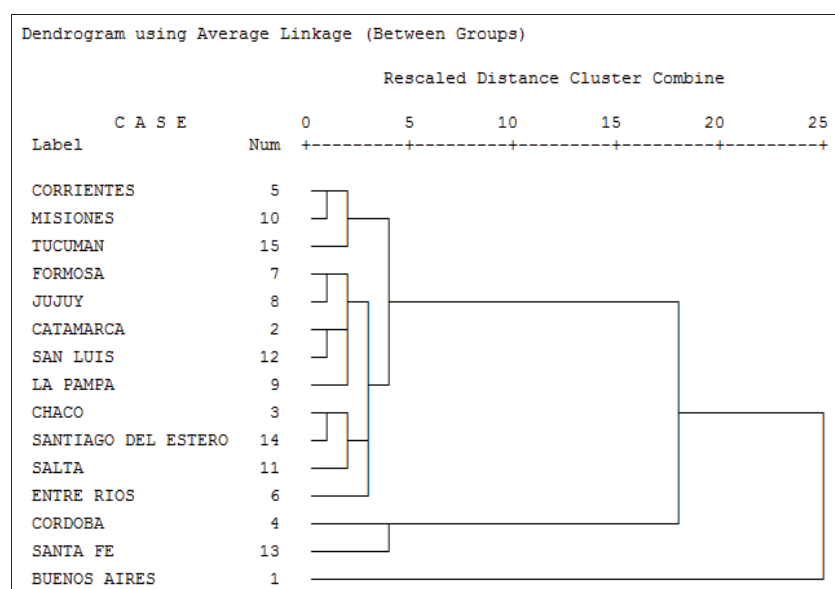


Figure 4: Dendrogram for the factors population and productive.

## 4 Conclusions

Soya cultivation begins in Argentina in the 1970s as an economic alternative to feed animals. Until the 1980s the cultivated area increases at a slowly rate, being these years when the expansion starts. The efficiency in soy cultivation has varied during the years without a defined pattern in the average, median, minimum and maximum values. Soya production depends on harvested and seeded areas in a direct form. There is a positive correlation between these variables that can be described and quantified with a linear regression model. Also, it has been possible to group the regions with similar characteristics when considering the variables related to soy production.





## References

- [1] Pengue, W. (2001). "Expansión de la soja en Argentina. Globalización, Desarrollo Agropecuario e Ingeniería Genética: Un modelo para armar", Buenos Aires, 2001. pp. 11. <http://www.grain.org/article/entries/453-expansion-de-la-soja-en-argentina> [enero 2013].
- [2] MinAgri (2012) <http://www.siaa.gov.ar/index.php/series-por-tema/agricultura> [abril 2012].
- [3] Falasca, S.L.; Miranda del Fresno, M.C. and Ulberich, A.C. (2012). "Potenciales consecuencias ambientales del fenómeno de sojización en Argentina". Congreso Argentina Ambiental 2012, I Congreso Internacional de Ciencia y Tecnología Ambiental y I Congreso Nacional de la Sociedad Argentina de Ciencia y Tecnología Ambiental. Mar del Plata. 2012. ISBN 978-987-28123-2-4, pp. 272-277.
- [4] Teubal, M. "Expansión del modelo sojero en la argentina, de la producción de alimentos a los commodities". pp. 71-96. Revista Realidad Económica N° 220.
- [5] Dros, J.M. (2004) Manejo del boom de la soya: Dos escenarios sobre la expansión de la producción de la soya en América del Sur. AID Environment. Amsterdam, junio 2004. [http://assets.panda.org/downloads/managingthesoyboomspanish\\_57b6.pdf](http://assets.panda.org/downloads/managingthesoyboomspanish_57b6.pdf).
- [6] Altieri, M. and Pengue, W. (2005), "La soja transgénica en América Latina: una maquinaria de hambre, deforestación y devastación socioecológica". Revista Ecología y Política N° 30. Ed. Icaria. Barcelona. pp. 87-93. <http://www.ecologiapolitica.info/ep/30.pdf>

